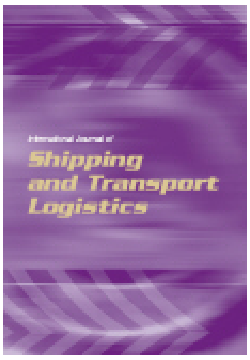


International Journal of Shipping and Transport Logistics

ISSN online: 1756-6525 - ISSN print: 1756-6517
https://www.inderscience.com/ijstl

Container transaction type prediction: a seaport case in Turkey

Elifcan Dursun, Sule Gungor

DOI: 10.1504/IJSTL.2022.10044799

Article History:

Received:	04 June 2021
Accepted:	01 December 2021
Published online:	07 August 2023

Container transaction type prediction: a seaport case in Turkey

Elifcan Dursun* and Sule Gungor

Tarsus University,
Vocational School,
International Trade,
Tarsus, Mersin, Turkey
Email: elifdursun@tarsus.edu.tr
Email: sulegungor@tarsus.edu.tr
*Corresponding author

Abstract: Container reshuffle is one of the main problems that container terminals face for several reasons. One reason for container reshuffle is uncertain transaction type. Yard planner needs the information for the transaction type to allocate inbound containers without causing a reshuffle. The vessel agent submits the transaction type information on the discharge list. However, before the vessel's arrival, circumstances – such as change of the cargo owner or lack of information – are encountered; therefore, information on the discharge list is unreliable. Yard planner can know the exact transaction type only before the container exits. This article follows the given steps of the cross industry standard process for data mining (CRISP-DM) at a seaport in Turkey to predict the transaction type before vessel arrival. We propose a multiple logistics regression model integrated with the terminal operating system to provide sustainable outputs to planners. The model predicts the container transaction type with 89% accuracy.

Keywords: container reshuffle; container transaction type; CRISP-DM; multiple logistics regression; seaport; terminal operating system.

Reference to this paper should be made as follows: Dursun, E. and Gungor, S. (2023) 'Container transaction type prediction: a seaport case in Turkey', *Int. J. Shipping and Transport Logistics*, Vol. 17, Nos. 1/2, pp.41–59.

Biographical notes: Elifcan Dursun is an industrial engineer and has eight years of port experience as senior process excellence specialist. Her work is focused on operations research and data science in logistics and supply chain management.

Sule Gungor is a Lecturer with four years of marketing experience at a seaport. Her research area is mainly focused on the supply chain economics.

1 Introduction

One of the main functions of seaports is container storage. Containers are stacked through various criteria to efficiently use the seaport's storage area, which sometimes introduces the container reshuffle problem along with storage. Container reshuffle occurs when the required container is in the lower tier at the stack, or the container is transferred to another area (Zeng et al., 2017). Terminal operations take one or more unproductive moves to reach the container, resulting in inefficiency at seaport operations (Bisira and Salhi, 2021).

Multiple reasons cause unproductive moves. Stacking containers in numerous tiers, vessel delays, vessel and container renomination, human error, and misinformation contribute to unproductive moves (Zhou et al., 2020).

One of the misinformation types encountered at seaports during the yard planning process is uncertain transaction type. Transaction type indicates what kind of transaction the container will encounter. In many seaports, containers are handled as inbound, outbound, or transshipment (Wang et al., 2014). Inbound containers are generally subject to an IM trading process. However, after the inbound containers are discharged in some seaports, they are not just transacted through the IM trading process, as they are sent to a third country by road. This type of transport is called TR transport (Aldis and Skapars, 2013).

This study focuses on the misinformation related to the uncertain transaction type of inbound containers that causes reshuffling. We consider misinformation to be inaccurate or incomplete information sharing from different sources such as vessel agents. Data is highly significant for many operations, including stacking, in seaports because it provides input for seaport operations. In our case study, seaport operations suffer from vessel agents' inaccurate and incomplete data submission resulting in poor yard planning and container reshuffling in the stacking yard.

Our case study has two different transaction types for inbound containers: import (IM) and transit (TR). The customs procedures of these two transaction types differ; thus, yard planners should not stack them together. TR transaction type requires different documentation processes and a vehicle tracking device for customs to monitor the truck on its way to the third country by road than IM transaction type, which affects the dwell time of the TR container compared to IM type. The vessel agent and yard planner are unaware of the latest consignee and the transaction type until the customer demands the container exit. Therefore, when submitting the discharge list to the seaport, the vessel agent specifies the transaction type based on the customer declaration. Nevertheless, due to human error reasons, misinformation is present in the transaction type. Yard planner uses the transaction type data provided to the TOS for inbound container stacking. Given the stacking plan based on the inaccurate data, seaport operations face container reshuffling in the yard.

Our study aims to reduce uncertain transaction type and reshuffling caused by misinformation at the seaport operating in Turkey. The CRISP-DM, one of the most critical process models, was applied to achieve this aim. During the modelling phase of the CRISP-DM, we developed a prediction model predicting the container's transaction type (IM/TR) on the discharge list before the vessel arrival, where the yard planner plan the inbound containers accordingly. After successful results, the prediction model was integrated with the TOS to provide sustainable outputs to yard planners.

In the subsequent parts of the study, we focused on the CRISP-DM, data collection, implementation steps of the CRISP-DM, proposed prediction model, accuracy, deployment, and recommendations.

As data science became more prominent globally than before, seaports began to pay attention to the field to utilise it and make more accurate predictions and plans (Lin et al., 2019). Many seaports shaped their organisational structures in this direction; likewise, the subject seaport plans to establish a data analytics unit. Seaport employees learned the methodology of data analytics through the application as a contribution of our study. In addition, they adopted processes to manage misinformation. Another contribution is that the yard planners can automatically reach the prediction model results through integration with TOS. The study is an example for other seaports that encounter different transaction types in inbound containers, such as seaports in Turkey, to be aware of uncertain transaction-induced reshuffling. Other seaports experiencing the same problem may easily apply our study to reduce the shuffling caused by uncertain transaction types.

2 Data-driven approaches for reshuffle problem

This section included the data mining, the CRISP-DM, and the data-driven approaches used to reshuffle problems encountered in ports. From the literature review we performed, it can be seen that literature on the CRISP-DM-oriented reshuffle problem performed at ports is lacking. Therefore, our paper attempts to fill a gap in the maritime industry and literature by adopting a data mining process approach to reduce yard reshuffling.

We focused on the CRISP-DM approach because the subject seaport wants to create a data analytics culture and place data mining, an essential part of data analytics, into this culture. In addition, our case has complex processes such as deploying the model in TOS. Therefore, a process-based data mining approach fits our case study.

Many different sectors, including maritime transport and seaport, have been using data mining to solve data-driven problems. In addition, data mining has become a significant concept for different industries due to utilising large amounts of data for prediction and analysis (Han and Kamber, 2016).

Data mining is not solely limited to extracting and sorting out the data. Data mining is a process to understand the patterns within the dataset to predict the desired output (Kotu and Deshpande, 2015). Therefore, the CRISP-DM presents a holistic process model to the analyst.

The CRISP-DM is a structured approach to planning a data mining project. It consists of business understanding, data understanding, data preparation, modelling, evaluation, and deployment phases. (Cazacu and Titan, 2020).

Although data science has become the leading term over data mining, it still preserves its standards on different projects. Martinez-Plumed et al. (2021) suggested that the CRISP-DM still fits the purpose for data science projects with its capability to promote different models during the modelling stage.

The reshuffle problem, which we handle as a data mining process, is seen as a problem most seaports face. Caserte et al. (2011) defined the reshuffling problem as the movement of the container within the yard. Typically, yard movements are a natural part of seaport operations. However, for many reasons, such as inaccurate information, containers are retrieved when storing them in some cases (Lee and Lee, 2010).

The way researchers deal with reshuffle problems differs. Like our study, Westbroek (2012) considered the display of yard movements caused by the change of container information as a holistic business intelligence application. He stated that it is necessary to examine the data mining phase, which is necessary for transforming data into information, with separate processes within business intelligence. On the contrary, some researchers see data mining as an essential step needed to prepare the data necessary for solving the problem, rather than presenting a process approach to data mining. For example, Gharehgozli et al. (2017) emphasised that vessel delay information should be obtained from many different terminal operators to estimate the yard reshuffle created by vessel delays, and this is a complex data mining study that needs to be studied with different databases. However, instead of applying a process approach for a complex data mining problem, they focused on the estimation model for problem-solving.

The starting point of our study is container reshuffling originating from the uncertain transaction type. However, the reason for the reshuffling of each seaport may not be the same. Zhou et al. (2020) characterised reshuffling because of the space allocation problem. They focused on data mining as a separate process to create information for the mixed-integer linear programming model to assign container locations to minimise yard reshuffle. Kourounioti et al. (2016) provide a different approach on reshuffles related to the space allocation problem. They predicted the dwell time of the IM containers to control the reshuffling related to the space allocation problem. Moreover, they developed the prediction model using the artificial neural networks approach during the modelling stage of data mining.

A reshuffle may occur in the yard caused by inaccurate data as well. For effective container stacking and to prevent reshuffling, the weight information of the containers must be accurate. Kang et al. (2006) developed a simulated annealing algorithm against inaccurate weight information-related reshuffling during the modelling phase of data mining. Apart from the modelling phase, they also indicated the importance of data cleaning within the scope of data mining.

Data mining has become an inevitable process with the increasing need for the big data approach in container terminals. Although Novaes Mathias et al. (2019) did not directly handle the reshuffling problem, they provide a data-driven approach for a waste problem correlated with container reshuffling. They highlighted the importance of data mining for the projects run in container terminals to describe the actual behaviour of any operation. They focused on a big data analytics framework to evaluate the terminal operations and address possible resource waste. As well as identifying the resource waste within the terminal, they found the erroneous data thanks to the data mining stage of the study.

3 Research methodology

The aim is to predict the transaction type before vessel arrival to reduce the reshuffle at the seaport, where we carried out the case study. Predicting the transaction type is essential for our study since we cannot reduce the reshuffling without identifying the container transaction type in the discharge list. The seaport has previously negotiated with the vessel agencies to specify a correct and precise transaction type in the discharge list. Nevertheless, due to reasons such as the latest consignee and the customer's erroneous declaration, agents cannot control the quality of the information. Therefore, a prediction model is needed in the study.

Data quality and data analysis are crucial for the prediction model. Therefore, we adopted a process approach and applied the steps of the CRISP-DM to the entire case study. The CRISP-DM is preferred in many projects as it offers a detailed description of stages, tasks and activities, and an actual data mining methodology (Palacios et al., 2017). In addition, complex enterprises such as seaports perform data mining based on specific methodologies and ensure project sustainability. The CRISP-DM steps followed in the study were:

1 business understanding

2 data understanding

3 data preparation

4 modelling

5 evaluation

6 deployment.

3.1 Data collection

We applied the CRISP-DM steps at the port, among the top five seaports in Turkey's container handling volume. The seaport offers conventional cargo and dry port services as well as container handling services. We did not mention the seaport information in the study due to corporate privacy reasons. For data analysis, we used three different tables in the database for 2019. These tables include container id, no, weight, size, type, category, status, regime, cargo description, operation type, loading port, and other relevant data.

Different employees from various units took part in the case study. Four people participated in the study: two from the data analysis unit, one from the information technology unit, and one from the operations unit.

3.2 Business understanding

In this part, we evaluated the causes of the problem that led to the case study. The case study aims to build a prediction model to predict the inbound container transaction type before vessel arrival and support the reduction of container reshuffle. The output of the prediction model is 1/0, where one indicates the TR container and 0 indicates the IM container.

Although this step of the study may seem trivial, it is crucial for data mining projects. It is impossible to implement the later steps without fully understanding the problem. We undertook the business understanding step by organising week-long meetings with the seaport staff participating in the case study. Depending on the nature of the problem, the duration of business understanding may vary between seaports.

3.3 Data understanding

Data understanding is the second step of the process model. In the dataset for 2019, inbound containers consisted of 26% TR and 74% IM containers. According to the information provided by the port participants, overall data accuracy for TR containers was 17% in 2019. We studied the dataset from three tables in the database during this process. We used the R Studio and Microsoft BI to work with the data in the tables. We merged the tables and finalised them for use in the prediction model. We completed the data understanding process in two weeks. The latest data frame includes the below variables:

- weight (numeric)
- size (factor)
- height (factor)
- category (factor)
- type (factor)
- status (factor)
- Pol (factor)
- ope`rator (factor)
- cargo description (factor)
- consignee (factor)
- dwell days (numeric) independent variables
- regime/transaction type (factor) (TR:1 IM 0) dependent variable.

3.4 Data preparation

In the third step, we prepared the data included in the prediction model to provide input. In our case study, data preparation consists of three steps. These steps are examination, reclassification, and transformation of the data. We completed the data preparation process in three weeks.

3.4.1 Data examination

We primarily used the Microsoft BI and R. During the data understanding step, we created the model outlook, which contained all the variables to predict the type of operation. We observed that vessel agents used different characters, misleading punctuation marks while providing the operator, consignee, and cargo description data. Therefore, we did not show the relationship between operator, consignee, cargo description, and transaction type. We emphasised that a separate data cleaning study should be carried out that includes the port staff involved in the study and vessel agents.

3.4.1.1 Size and transaction type

According to the analysis, 79% of inbound containers with TR transaction type are 40 foot (40’) containers. We have seen that the size variable is not a determinant for the IM transaction type. We showed the relationship between size and transaction type in Figure 1.

3.4.1.2 Height and transaction type

Commonly, the height of a standard container is 8.6 feet (806’), whereas a HC container height is 9.6 feet (906’). 9 feet (900’) containers are not standard in the shipping industry, but they are often requested for storage to fill a specifically sized gap (World Shipping Council, 2014). When we investigated the relationship between height and transaction type, many (74%) containers traded with TR transaction types were also HC containers. For the IM transaction type, we could not mention a majority. We showed the relationship between height and transaction type in Figure 2.

Figure 1 Size and transaction type relationship

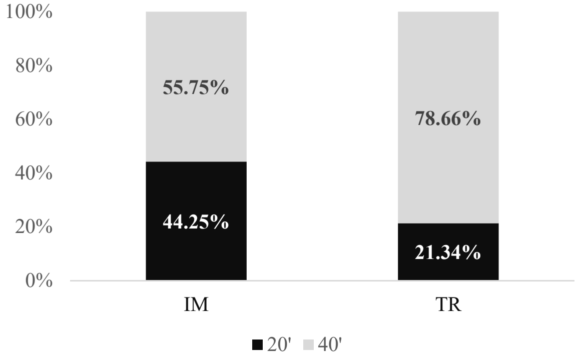


Figure 2 Height and transaction type relationship

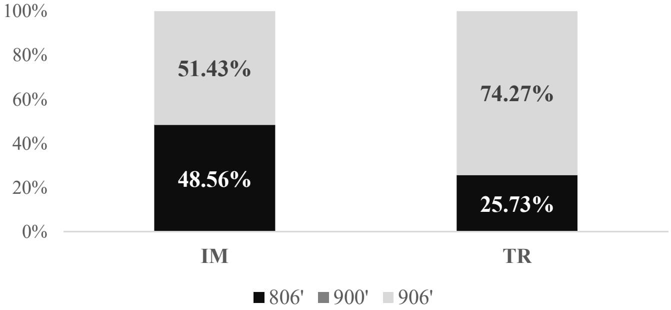
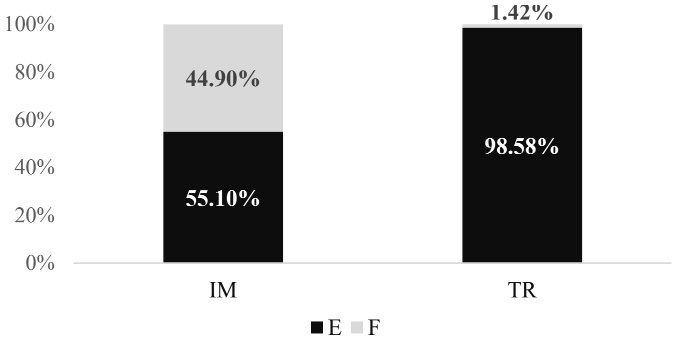


Figure 3 Status and transaction type relationship



3.4.1.3 Status and transaction type

Container status means that the containers are full or empty. In the study, E represents empty containers, while F full containers. 99% of the containers traded with TR transaction type are discharged as laden. We could not make a decisive distinction for the IM containers. We showed the relationship between status and transaction type in Figure 3.

3.4.2 Data reclassification

In the data understanding step, some variables – such as the Pol, category, and cargo description – had too many factors that the model could not operate. In this step, we sought to reclassify each of these variables in meaningful ways. Therefore, we have re-examined all variables with a high factor value of more than twenty and classified them for modelling.

3.4.2.1 Category reclassification

There were 21 unique category factors in our data framework. However, only seven categories accounted for 99% of the total volume in the dataset. The remaining 1% was classified as 'Others,' creating a total of eight-factor category variables.

3.4.2.2 Type reclassification

We applied the same approach for the type variable. We observed seventeen different container-type factors in the data frame. However, only four of them accounted for 99% of the total volume. The remaining 1% were classified as 'Others.' Therefore, we created a new type variable consisting of five factors. In the model, we used the new type variable.

Table 1 Reclassification of Pol

Transit status	*Frequency*	*New Pol classification*
0.0–0.1	34	10
0.1–0.2	9	9
0.2–0.3	15	8
0.3–0.4	12	7
0.4–0.5	4	6
0.5–0.6	2	5
0.6–0.7	2	4
0.7–0.8	3	3
0.8–0.9	3	2
0.9–1.0	6	1

3.4.2.3 Port of loading (Pol) reclassification

Pol refers to the port where cargoes or containers are loaded onto a vessel. We observed 288 different Pol factors in the data frame. First, we grouped each Pol according to its country. Thus, we reduced the number of factors to 90. Nevertheless, that figure was still too much to run on the model. Therefore, we examined the TR status of each loading port by country. The TR status aimed to determine how many TR containers discharged from each port Pol allocated on a country-by-country basis in the data frame. According to the value of TR status, we reclassified the countries and divided them into ten different factors. For example, an examination of the Pol of inbound containers by Belgium (BE) country code showed that 1,790 containers were discharged in 2019. TR containers accounted for 34% of the total containers for BE country code. Following the logic of reclassifying, shown in Table 1, the TR status corresponded to Group 7. Therefore, BE country-code Pol was reclassified to be in group 7 in the study.

We performed the same operation for all Pol country groups, reducing factors to 10. We used the new Pol classification as input in the model.

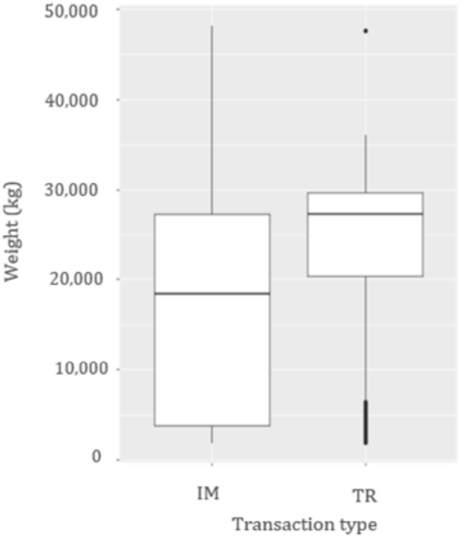
3.4.3 Transformation

In this step, we analysed whether the variables fit the normal distribution before running the model. We transformed the variables that do not follow normal distribution to fit the normal distribution. We analysed the weight and the dwell time-continuous variables in the normal distribution.

3.4.3.1 Weight and transaction type

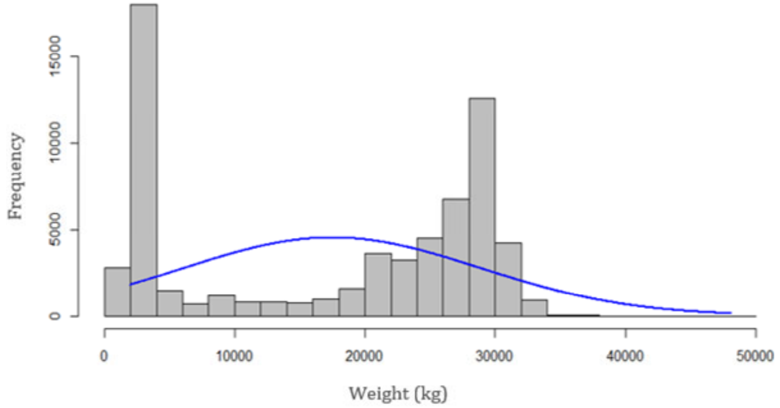
We used the box plot graph to compare the weight distributions of TR and IM containers. Researchers generally use the box plots to visualise and compare groups of data (Kendrick, 1989). When we examined the relationship between weight and transaction type with the median values of the box chart, it appeared that IM and TR differ. Therefore, the weight of inbound containers can be decisive in transaction type prediction. Figure 4 shows the box plot graph of transaction type and weight.

Figure 4 Transaction type and weight box plot



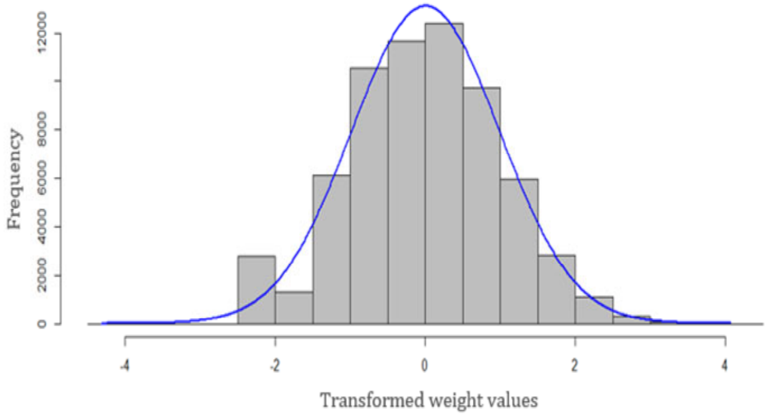
A linear relationship between the dependent and independent variables is not required (Schreiber-Gregory, 2018) in logistics regression. Although there is no such requirement, we applied a normality test for the weight variable in our case study. We understood that the weight variable does not correspond to the normal distribution. Figure 5 shows the non-normal distribution of the weight variable.

Figure 5 Non-normal distribution of weight (see online version for colours)



We transformed the non-normal distributed weight variable into a format corresponding to normal distribution considering further studies. In the absence of connections, ordered quantile normalisation (ORQ) can generate normally distributed transformed data (Peterson and Cavanaugh, 2019). In addition, the ORQ uses the original values of a sample to estimate a normalising transformation function with a semiparametric approach. For this reason, we used ORQ in the study to transform the weight variable suitable for normal distribution. Figure 6 shows the normal distribution of the weight variable after transformation.

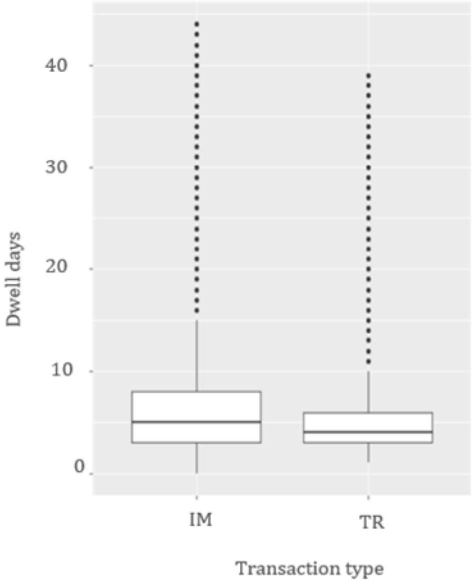
Figure 6 Normal distribution of transformed weight (see online version for colours)



3.4.3.2 Dwell days and transaction type

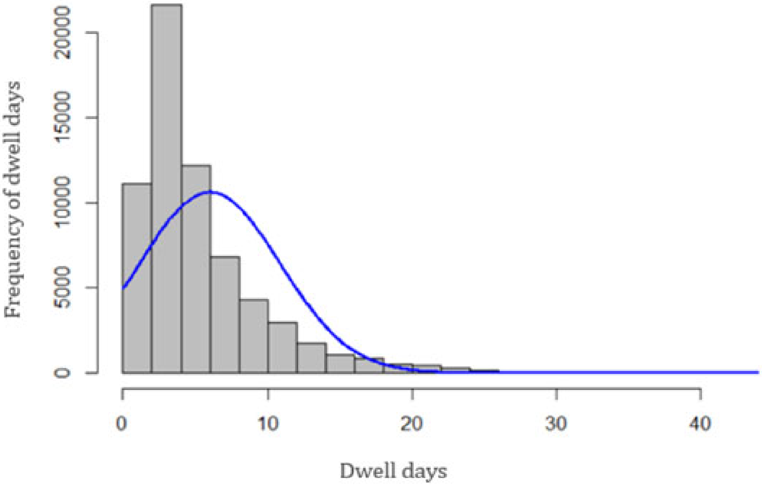
When we examined the relationship between dwell days and transaction type in the box chart, we understood no significant difference between IM and TR transaction types. Outliers were present in both TR and IM transaction type, as shown in Figure 7.

Figure 7 Transaction type and dwell days box plot



When we tested the dwell days variable for normality, it became clear that the values did not correspond to the normal distribution. Figure 8 shows the non-normal distribution of dwell time.

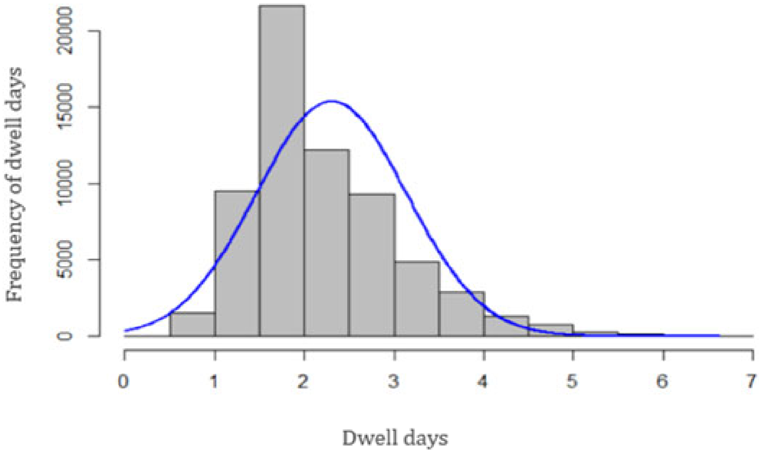
Figure 8 Non-normal distribution of dwell days (see online version for colours)



When the residuals of a response variable are not normally distributed, the response variable is suitable for the normal distribution by using specific transformation approaches. The aim is to transform the response variable from y to $\sqrt{y}$ while applying square root transformation (Gregoire et al., 2008). Applying zero values is also considered an advantage of square root transformation (Manikandan, 2010). For this reason, we applied the square root transformation in the study for the dwell days variable.

Figure 9 shows the normal distribution of dwell days after the transformation process. We used the square root transformation to fit the normal distribution.

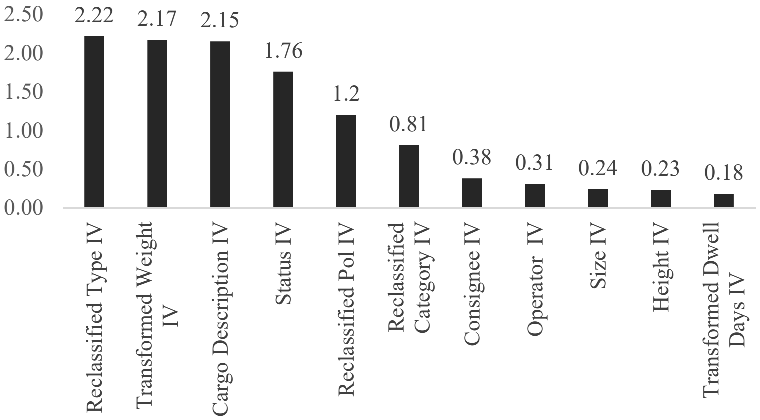
Figure 9 Normal distribution of dwell days after transformation (see online version for colours)



3.4.4 Importance level

We reviewed all independent variables and their relationships with the transaction type. The critical point to look out for in this section was whether each variable was essential for predicting the transaction type. In order to answer this, we applied relative importance analysis in R Studio. Relative importance refers to the contribution a variable makes to predicting a dependent variable by itself and in combination with other predictor variables (Johnson and LeBreton, 2004). As a result of the analysis, we have transpired that type, weight, cargo description, and status were the most important variables to predict the transaction type. Figure 10 shows the relative weights of independent variables.

Figure 10 Importance levels of independent variables (IV)



3.5 Modelling

After data preparation, we developed a multiple logistics regression model for the transaction type prediction of inbound containers. Logistics regression is an accepted method in probability modelling the binary dependent variable (Yalcin et al., 2011). In the study, we used logistics regression, as it takes binary values in the case of transaction type TR : 1 IM : 0, which is the dependent variable. According to the correlation matrix output, we observed that transformed weight and status variables are highly correlated (r = 0.79). Therefore, we included the status variable to avoid the multicollinearity problem in our model. In addition, we used reclassified type due to its importance and the reclassified Pol to predict the transaction type-dependent variable. We did not include cargo description since it requires a separate, sustainable data cleaning study from port and vessel agents. We apportioned the data into training and test sets with a 70%–30% split. The entire modelling step lasted one week.

The multiple logistics regression model developed was:

$$Y = Logit(p) = ln\left(\frac{p}{1-p}\right) \tag{1}$$

$$Y = C_0 + C_1 * X_1 + C_2 * X_2 + \cdots + C_n * X_n \tag{2}$$

$$Transaction\ Type(1,0) = C_0 + C_1 * Type + C_2 * Port\ of\ Loading + C_3 * Status \tag{3}$$

P represents the probability that the dependent variable (*Y*) is 1. *p*/(1−*p*) is the so-called odd or frequency ratio. C_0 is the intercept, and C_1, C_2,…, C_n are coefficients. Coefficients measure the contribution of the independent factors to the variations in *Y* (Lee, 2005).

Table 2 Coefficient table

	Estimate	*Std. error*	*z value*	*Pr (> |z|)*
(Intercept)	–0.59012	0.20910	–2,822	0.00477**
New_Type_ColumnHC	0.35691	0.03597	9,922	< 2e-16***
New_Type_ColumnOT	0.32869	0.24709	1,330	0.18344
New_Type_ColumnOther	0.53958	0.13289	4,060	4.90e-05***
New_Type_ColumnRH	2.01162	0.06457	31,155	<2e-16***
Status_F	3.33621	0.08638	38,621	<2e-16***
New_Grouped_POL_Country_10	–3.79576	0.21074	–18,012	<2e-16***
New_Grouped_POL_Country_2	–0.06320	0.75451	–0.084	0.93325
New_Grouped_POL_Country_3	–1.31979	0.29433	–4,484	7.33e-06***
New_Grouped_POL_Country_7	–2.52268	0.20434	–12,345	<2e-16***
New_Grouped_POL_Country_8	–2.79018	0.20363	–13,702	<2e-16***
New_Grouped_POL_Country_9	–3.40174	0.21228	–16,025	<2e-16***
New_Grouped_POL_Country_1	–1.33750	0.25563	–5,232	1.68e-07***
New_Grouped_POL_Country_4	–1.45971	0.80951	–1,803	0.07136
New_Grouped_POL_Country_5	–1.93660	0.40331	–4,802	1.57e-06***
New_Grouped_POL_Country_6	–1.61574	0.33129	4,877	1.08e-09***

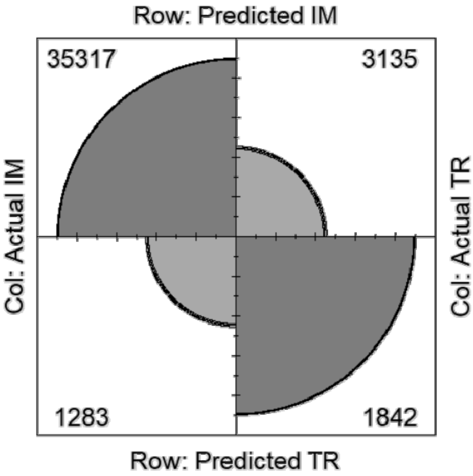
We developed the multiple logistic regression model in R Studio. We presented the coefficient values for the model's output in Table 2.

The intercept and the coefficient estimate correspond to each predictor variable. For instance, the coefficient estimate of the variable type of HC is b = 0.3569, which is positive. This value refers to an HC container associated with an increase in the probability of a TR container. For every unit change in HC type, the log odds of TR container increases by 0.3569. The standard error (std. error) of the coefficient estimates represents the accuracy of the coefficients. A more significant standard error indicates less confidence in the estimate (Sperandai, 2014). The p-value (*Pr* (>|*z*|)) indicates how significant the model estimate is. A smaller p-value indicates that the estimate is more significant (Menard, 2011). In our case study, most of the factors used as input in the model are highly significant to predict the transaction type.

3.6 Evaluation

We measured the validity of the generated multiple logistic regression model. We plotted a receiver-operating characteristic (ROC) curve to show the trade-off between sensitivity and specificity. The area under the ROC curve (AUC) value lies between 0.5 to 1, where 0.5 denotes a poor classifier, and 1 denotes an excellent classifier (Zou et al., 2007). In our case, we measured the AUC value as 0.84, which is considered excellent. Sensitivity explains the percentage of TR containers predicted correctly. The model predicted 96% of IM containers and 37% of TR containers correctly. We measured the overall model accuracy as 89%, which is good and acceptable for our case. Figure 11 shows the actual/predicted number of IM and TR containers.

Figure 11 Confusion matrix



According to the confusion matrix, out of 36,600 IM containers, the model predicted 35,317 of them as IM containers, whereas predicted the 1,842 TR containers out of 4,977 TR containers.

We analysed each Pol country factor to obtain a more accurate prediction result, although the model's accuracy was acceptable. We understood that the most erroneous prediction outputs were the containers with the Egypt Pol mark.

Table 3 Incorrect outputs based on Pol

Pol country	*Accuracy %*	*Inaccuracy %*	*Number of incorrect prediction*
EG	89%	11%	1,127
CN	80%	20%	542
NL	43%	57%	274
ES	92%	8%	243
BE	79%	21%	228
MT	86%	14%	186
US	86%	14%	153
DE	76%	24%	137
IN	72%	28%	134
SA	91%	9%	113
Others	91%	9%	1,281
Total			4,418

Table 4 Inbound containers with Egypt Pol mark

No	*REJIME_C <fct>*	*New_Type_C <chr>*	*Status_C <chr>*	*Total container <int>*
1	0	DC	E	772
2	0	DC	F	315
3	0	HC	E	554
4	0	HC	F	981
5	0	OT	E	47
6	0	OT	F	4
7	0	Other	E	19
8	0	Other	F	23
9	0	RH	F	124
10	1	DC	E	26
11	1	DC	F	460
12	1	HC	E	6
13	1	HC	F	1,051
14	1	OT	F	7
15	1	Other	F	52
16	1	RH	F	1,419

We sought the problem in containers arriving with the Egypt Pol mark. The model successfully predicted the container's transaction type with reefer high cube (RC) type, full status, Egypt Pol mark. It is because full RH-type containers are decisive for the TR transaction type. However, the same does not apply for the full, dry, and high cube

(DC/HC) container type with Egypt Pol mark. For this reason, the model may not accurately predict the transaction type of these containers. Table 3 shows the distribution of 4,418 containers incorrectly predicted on a Pol country basis. Table 4 shows the details of inbound containers with Egypt Pol mark.

We presented the model outputs and evaluated them together with the yard planning unit. The evaluation process lasted a week with a discussion of the next steps.

3.7 Deployment

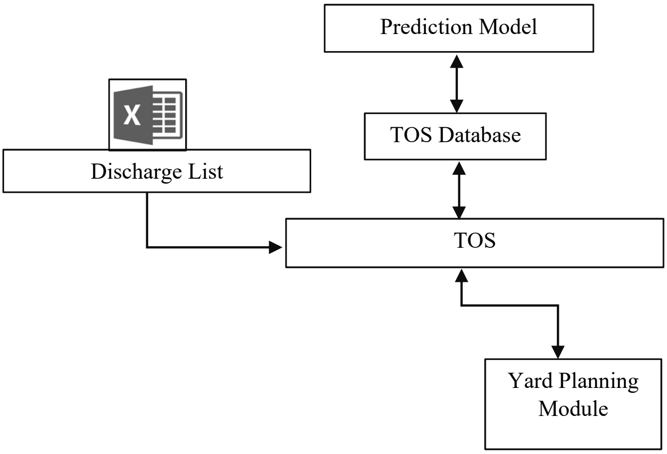
In the final step of the CRISP-DM method, the multiple logistics regression model was integrated with the TOS used in the port. Integration was essential for yard planners to use the model dynamically. Figure 12 shows the deployment that takes place on TOS.

The model is automatically operated every 4 hours. Thus, when the vessel agent submits the discharge list to TOS, the model makes a transaction type prediction using the data on the discharge list. Prediction results are dynamically transferred to the yard planning module through TOS deployment. Negotiations, detailed deployment analysis, and deployment stages with the port's TOS vendor company were completed in two and a half months. On the planning screen, yard planners display the data on the discharge list and the prediction output as a summary in Table 5.

Table 5 Summary output of the deployment

CNTR_N	*BL_N*	*REJIME_C*
Container number	Bill of lading number	Transaction type
Transit:1		
Import: 0		

Figure 12 Model deployment



Yard planners allocate the inbound containers based on the summary output. In addition, TOS vendor created a reason section on the yard operator screen to understand the impact of the study on reshuffling. The reason section enables the yard operator to select the reshuffling reason. Reason selection is of great importance in monitoring the seaport's yard reshuffling and, therefore, its operational performance. On the other hand, selection has a disadvantage, such as an operator selecting the transfer type incorrectly. The generated report eliminated this disadvantage by examining the success of the model deployment. The report shows the final and source destinations of the reshufflings carried out by the yard operator. Although the yard operator selects the uncertain transaction type as the reason for reshuffling, a transfer to an area rather than the TR area proves to be an incorrect selection. The erroneous reshuffling reason selection rate was 2% at the seaport, which was acceptable for our case study.

Yard planners initiated the usage of the deployed output in 2020 November. They have been planning the inbound containers accordingly. We received a 4% decrease in overall yard reshuffling in 2021 January compared to 2020 December.

4 Conclusions

Reshuffling, which takes place in the yard and affects the efficiency of port operations, is the act of changing a blocking container to another location (Tang et al., 2014). Container reshuffling occurs for several reasons. One of these reasons is the uncertainty of the transaction type of inbound containers. In the seaport where we carried out our case study, inbound containers have two different transaction types: TR and IM.

TR and IM transaction types are different from each other in customs processes. Therefore, their stacking together causes container reshuffle. We developed a multiple logistics regression model that predicts the transaction type before vessel arrival to reduce container reshuffle caused by uncertain transaction type. We continued all the steps within the framework of the CRISP-DM methodology. In our case study, we used the CRISP-DM to lay a ground for data analytics culture at the seaport.

We used the reclassified type, Pol, and status variables in the multiple logistics regression model. The model predicts the IM and TR inbound containers with 89% accuracy. Overall data accuracy for TR containers was 17% in 2019, whereas the model predicted 37% of the TR containers accurately.

A detailed examination of the incorrect prediction results showed that most were inbound containers with the Egypt Pol mark. The main reason for incorrect prediction is that the type variable was decisive in distinguishing between TR , IM , and lost importance in containers with the Egypt Pol mark. The prediction model was therefore unable to make an accurate distinction. At this stage, we recommended testing the accuracy rate by adding different variables to the model. In the case study, port operations managers determined that the accuracy rate was sufficient for the model deployment. The model was integrated with the TOS at the final step, and the results were automatically transferred to the yard planning module.

With the help of the model, yard planners can now plan inbound containers by observing the transaction type separation on the discharge list before the vessel arrives. In addition, the TOS vendor created a transfer reason section on the yard equipment operator screen to monitor container reshuffles. As a result, yard operators can select the reshuffling reason as an uncertain transaction type. We received a 4% decrease in overall yard reshuffling in 2021 January compared to 2020 December after yard planners initiated to implement the deployed output in 2020 November.

Our study sets an example for other ports with uncertain transaction type problems. In addition, when we examined the previous academic studies, we did not find a process-based data mining study applied in ports. Our study approached the prediction of transaction type as a data mining process that aims to reduce reshuffles eventually as an improvement presented by the prediction model. Moreover, it contributes to academic studies and industry with the model developed to predict uncertainty as a step of the data mining process.

Acknowledgements

The open access fee of this work is supported by China Merchants Energy Shipping.

References

Aldis, B. and Skapars, R. (2013) 'Development of international freight transit in Latvia', *Procedia – Social and Behavioral Sciences*, Vol. 99, No. 1, pp.57–64.

Bisira, H. and Salhi, A. (2021) 'Reshuffle minimisation to improve storage yard operations efficiency', *Journal of Algorithms and Computational Technology*, Vol. 15, No. 1, pp.1–11.

Caserta, M., Schawarze, S. and Voss, S. (2011) 'Container rehandling at maritime container terminals', in *Handbook of Terminal Planning Operations, Research/Computer Science Interfaces Series*, Vol. 49, pp.247–269, Springer, New York, NY.

Cazacu, M. and Tıtan, E. (2020) 'Adapting CRISP-DM for social sciences', *Broad Research in Artificial Intelligence and Neuroscience*, Vol. 11, No. 2, pp.99–106.

Gharehgozli, A., Mileski, J. and Duru, O. (2017) 'Heuristic estimation of container stacking and reshuffling operations under the containership delay factor and mega-ship challenge', *Maritime Policy and Management*, Vol. 44, No. 3, pp.373–391.

Gregoire, T., Lin, Q., Boudreau, J. and Nelson, R. (2008) 'Regression estimation following the square root transformation of the response', *Forest Science*, Vol. 54, No. 6, pp.597–606.

Han, J. and Kamber, M. (2016) *Data Mining: Concepts and Technique*s, Second ed., Morgan Kaufmann Publishers, San Francisco.

Johnson, J.W. and LeBreton, J.M. (2004) 'History and use of relative importance indices in organizational research', *Organizational Research Methods*, Vol. 7, No. 1, pp.238–257.

Kang, J., Ryu, K. and Kim, K. (2006) 'Determination of storage locations for incoming containers of uncertain weight', in *Proceedings of International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, Berlin, pp.1159–1168.

Kendrick, J. (1989) 'The box plot: a simple visual method to interpret data', *Annals of Internal Medicine*, Vol. 110, No. 11, pp.916–921.

Kotu, V. and Deshpande, B. (2015) *Predictive Analysis and Data Mining: Concepts and Practices with Rapid Miner*, First ed., Morgan Kaufmann Publishers, San Francisco.

Kourounioti, I., Polydoropoulou, A. and Tsiklidis, C. (2016) 'Development of models predicting dwell time of import containers in port container terminals-an artificial neural networks application', *Transportation Research Procedia*, Vol. 14, No. 1, pp.243–252.

Lee, S. (2005) 'Application of logistics regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data', *International Journal of Remote Sensing*, Vol. 26, No. 7, pp.1477–1491.

Lee, Y. and Lee, Y. (2010) 'A heuristic for retrieving containers from yard', *Computers and Operations Research*, Vol. 37, No. 6, pp.1139–1147.

Lin, M., Li, M., Hao, H. and Zhen, L. (2019) 'A method for estimating liner shipping time under uncertainty', *International Journal of Shipping and Transport Logistics*, Vol. 11, Nos. 2–3, pp.145–160.

Manikandan, S. (2010) 'Data transformation', *Journal of Pharmacology and Pharmaco therapeutics*, Vol. 1, No. 2, pp.126–127.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Ramirez-Quintana, M., and Flach, P. (2021) 'CRISP-DM twenty years later: From data mining processes to data science trajectories', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, No. 8, pp.3048–3061.

Menard, S. (2011) 'Standards for standardized logistics regression coefficients', *Social Forces*, Vol. 89, No. 4, pp.1409–1428.

Noveas Mathias, T., Shinoda, T., Hangga, P. and Inutsuka, H. (2019) 'Big data approach to identify the waste management of container terminal resources', Asian Transport Studies, Vol. 5 No.4, pp. 653-678.

Palacios, H., Toledo, R., Pantoja, G. and Navarro, A. (2017) 'A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change', *Advances in Science, Technology and Engineering Systems Journal*, Vol. 2, No. 3, pp.598–604.

Peterson, R. and Cavanaugh, J. (2019) 'Ordered quantile normalization: a semiparametric transformation built for cross-validation era', *Journal of Applied Statistics*, Vol. 47, No. 13, pp.1–16.

Schreiber-Gregory, D. (2018) 'Logistic and linear regression assumptions: violation recognition and control', in *Proceedings of the 26th Annual Southeast SAS Users Group Conference*, Williamsburg, VA.

Sperandai, S. (2014) 'Understanding logistics regression analysis', *Biochemia Medica*, Vol. 24, No. 1, pp.12–18.

Tang, L., Jiang, W., Liu, J. and Dong, Y. (2014) 'Research into container reshuffling and stacking problems in container terminal yards', *IIE Transactions*, Vol. 47, No. 7, pp.751–766.

Wang, L., Zhu, X. and Xie, Z. (2014) 'Storage space allocation of inbound container in railway container terminal', *Mathematical Problems in Engineering*, Vol. 14, No. 1, pp.1–10.

Westbroek, M. (2012) *A Conceptual Business Intelligence Framework for the Identification, Analysis and Visualization of Container Data Changes and its Impact on Yard Movement*, MSc thesis, Open University of Netherlands.

World Shipping Council (2014) 'Containers', [online] https://web.archive.org/web/20141008144227/http://www.worldshipping.org/about-the-industry/global-trade, (accessed 15 October 2021).

Yalcin, A., Reis, S., Aydinoglu, A.C and Yomralioglu, T. (2011) 'A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods for landslide susceptibility mapping in Trabzon', *CATENA*, Vol. 85, No. 3, pp.274–287.

Zeng, Q., Feng, Y. and Chen, Z. (2017) 'Optimizing berth allocation and storage space in direct transshipment at container terminals', *Maritime Economics and Logistics*, Vol. 19, No. 1, pp.474–503.

Zhou, C., Wang, W. and Li, H. (2020)' Container reshuffling considered space allocation problem in container terminals', *Transportation Research Part E: Logistics and Transportation Review*, Vol. 136, No. 1, pp.18–34.

Zou, K., O'Malley, J. and Mauri, L. (2007) 'Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models', *Circulation*, Vol. 115, No. 5, pp.654–657.